Annual Progress Report

Office of Naval Research Grant No. N00014-95-1-1163 (AASERT)

Professor Avideh Zakhor, Principal Investigator

Period covered by report:

June 1, 1996 - May 31, 1997

Summary of Activities for 6/1/1996 through 5/31/1997

Avideh Zakhor
231 Cory Hall
University Of California
Berkeley, CA 94720

1 Introduction

The problems of image sequence compression and new view synthesis have both received a lot of attention recently. In the former case, it is desired to compactly represent the original image set by exploiting redundancy and correlation. This issue is particularly important in applications of storage and transmission. In contrast, the goal of new view synthesis is to generate arbitrary viewpoints of a given scene primarily for visualization purposes. Notice that there exists a trade-off between representation size and the quality of the synthesized images: As more views of the scene are added to the representation, the image quality increases as does the representation size. Hence, an interesting problem is to consider both problems at once; that is, construct a compact representation which reconstructs the original images and synthesizes new views.

We propose two depth-based representations to address these problems. The first approach involves several so-called *reference frames* for which depth and intensity information are both defined. New views are generated by warping the reference intensity and depth data in a manner similar to view interpolation techniques [3, 6, 10, 1, 2, 9, 5]. The second proposed approach integrates all available information with respect to a single reference frame akin to layered representations [7, 14, 13, 15]. The representation then consists of a multivalued array of depth and intensity values which overcomes occlusions and redundancy [4]. These depth-based representations both assume the given image sequences arise from a static 3-D scene captured by a moving camera restricted to the x-y plane. Note that the exact motion of the camera is unknown *a priori* and will be estimated.

The report is organized as follows. Section 2 describes techniques for estimating dense depth from a set of neighboring frames and for warping this information to synthesize new views. In Section 3, these techniques are applied to the proposed approach of view interpolation from multiple reference frames. In Section 4, a multivalued representation is presented which condenses and redefines the available information with respect to a single reference frame. Finally, extensions and areas for future work are discussed in Section 5.

2 Depth Estimation and Synthesis

The basis for the proposed representations is computing depth information at reference locations. Given a particular frame, the goal is to derive an estimate of depth for every pixel location using neighboring frames. Dense depth information is desired because it provides the appropriate mapping for every point during synthesis. Depth estimation and synthesis techniques are described in the following sections.

2.1 Dense Depth Estimation: Pairwise

Given an image sequence, it seems intuitive to compute depth pairwise between the reference frame and each of its neighbors to generate local "depth maps". Since every frame is related by a planar translation, depth estimation can be accomplished by 1-D correspondence matching along the parallel epipolar lines. In [1], the l_2 norm of intensity error is minimized over possible depth values using adaptive neighborhoods $\mathcal{N}$:

$$\min_d \left\{ \sum_{(u,v)} \sum_{\in \mathcal{N}} \| I_k(u,v) - I_i(u',v') \|^2 \right\} \tag{1}$$

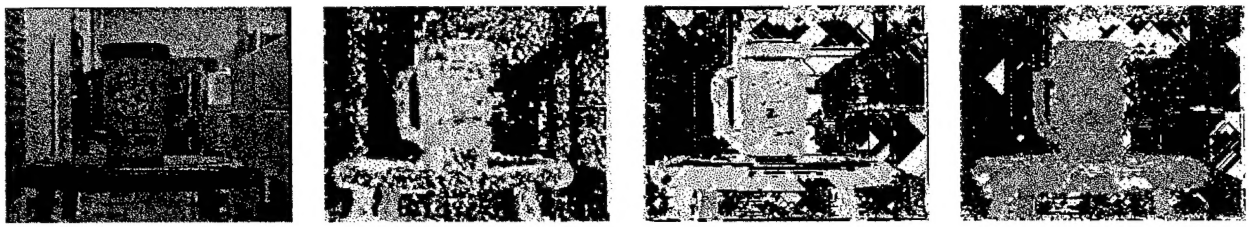
where predicted coordinates (u', v') and disparity d are related to a candidate motion vector (m, n) by

$$u' = u + m \tag{2}$$

$$v' = v + n \tag{3}$$

$$d = \sqrt{m^2 + n^2} = \frac{f}{z}\sqrt{b_x^2 + b_y^2} \tag{4}$$

Moreover, points which lead to potential artifacts and spurious matches are marked as low confidence. To improve the estimates, the local depth maps are normalized and then combined together to form a single estimate for the given reference frame. The combined result is more accurate because low confidence regions in the local depth maps do not always overlap, thus enabling higher confidence estimates to fill in these regions.



(a) (b) (c) (d)

Figure 1: *Example of depth maps using fixed and adaptive block sizes: (a) intensity reference frame; (b) matching with fixed* 9×9 *blocks; (c) matching with adaptive blocks; and (d) combining multiple local depth maps. The colors in (c) and (d) represent low confidence matches resulting from different artifacts as described in [1].*

As an example, consider Figures 1 (a)–(d). An intensity frame from the Mug2 sequence in Section 3 is shown in Figure 1 (a). To obtain dense depth information for this frame, one attempts to match it with each of its neighbors. Figure 1 (b) shows the resulting local depth map using a fixed 9×9 block size. While the mug and stool are somewhat discernible, there are a large number of artifacts throughout the scene due primarily to the many low-textured regions. In contrast, Figure 1 (c) shows a depth map obtained using an adaptive block size with various low confidence regions marked accordingly. Notice the improvement in depth estimation for the mug and stool as well as the background points. Figure 1 (d) demonstrates the effectiveness of combining several local depth maps together. The final depth map is a more accurate estimate of the given scene as compared with Figure 1 (c). The regions in the combined depth map which may be inaccurate are marked in yellow to indicate low confidence. To eliminate low confidence regions, one may apply a spline-based filling algorithm [1].

2.2 Dense Depth Estimation: Multiframe

While pairwise matching leads to reasonable depth results, multiframe approaches perform even better by reducing ambiguity and increasing accuracy when camera motion is known. To compute depth for a particular frame, a variant of Okutomi and Kanade's multiple-baseline algorithm is used [12]. The approach consists of finding the inverse depths that minimize the sum of component intensity errors. More precisely, suppose there are M images denoted by $I_i(\cdot,\cdot)$ and let $k \in 1, 2, \ldots M$ be the reference frame. Then, the goal is to compute inverse depth ζ for every desired point with the following expression

$$\min_{\zeta} \left\{ \sum_{i \neq k}^{M} \sigma_i \left(\sum_{(u,v)} \sum_{\in \mathcal{N}} \| I_k(u,v) - I_i(u',v') \|^2 \right) \right\} \tag{5}$$

where $\mathcal{N}$ is a local neighborhood around the pixel of interest, σ_i indicates the influence of frame i, and (u', v') are the predicted image coordinates. For planar translation, they are given by

$$u' = u - f b_{xi} \zeta \tag{6}$$

$$v' = v - f b_{yi} \zeta \tag{7}$$

Assuming the baselines (b_{xi}, b_{yi}) are known *a priori* or else computed, one can proceed to estimate the inverse depths ζ using Eqn (5) for all desired points in the frame.

Our implementation of the multiple-baseline algorithm differs from Okutomi and Kanade's in several ways. First, adaptive neighborhood sizes for $\mathcal{N}$ are employed to improve estimation in low-textured regions. The neighborhood is automatically adjusted according to the local variance of neighboring intensities [5]. Next, instead of normalizing the largest baseline to be 1, one of the shorter baselines is considered to have unity baseline. This feature permits wider baselines to be included without drastically increasing computational time.

Because wider baselines may be used, occlusions in the scene will pose a larger problem in multiframe matching. The effects of occlusions are mitigated by the addition of σ_i in Eqn (5) [4]. For a given point in the reference frame, it is very likely that no corresponding point is visible in at least one other frame of the sequence. Blindly including all frames in the minimization may lead to spurious results. One possibility is to examine the intensity errors associated with each frame i over the entire range of allowable ζs and then set $\sigma_i = 0$ for the point if the error exceeds some predetermined threshold. In this manner, only the frames in which the point can be seen contribute to the minimization.

The parameter σ_i may also reduce matching problems along edges. When edges in a pair of images are oriented in the same direction as the motion (assuming planar translation), typical algorithms produce spurious matches due to the limited extent of the local neighborhood, the so-called aperture ambiguity problem. A very reasonable solution would be to ignore frames whose edges are oriented parallel to the direction of camera motion.

2.3 View Synthesis

Once a dense depth map has been computed using either pairwise or multiframe matching, it is relatively straightforward to warp the reference information to synthesize new views of the scene. The procedure consists of regarding the depth map as a deformable mesh of quadrilateral patches [5]. Vertices of each patch are warped by the appropriate transformation. For reconstruction of the original images, the transformation is simply

$$u' = u + f b_x / Z \tag{8}$$

$$v' = v + f b_y / Z \tag{9}$$

where f is the focal length, (b_x, b_y) is the amount of planar translation, and Z is the depth corresponding to point (u, v). Alternatively, off-plane views may be obtained by using the transformation

$$u' = f\frac{r_{1,1}X + r_{1,2}Y + r_{1,3}Z + \Delta x}{r_{3,1}X + r_{3,2}Y + r_{3,3}Z + \Delta z} \quad (10)$$

$$v' = f\frac{r_{2,1}X + r_{2,2}Y + r_{2,3}Z + \Delta y}{r_{3,1}X + r_{3,2}Y + r_{3,3}Z + \Delta z} \quad (11)$$

The interior is rendered using a traditional 2-D scan-line algorithm and Z-buffering to ensure the proper depth ordering [8]. Patches which transcend depth edges are not rendered since they may lead to "smearing" [5]. In the end, it is possible for the final image to contain "holes" which correspond to slight inaccuracies in the estimated depth or to regions unseen in the original frames.

(a) (b) (c)

Figure 2: *Example of synthesizing new view from a single reference pair: (a) intensity image frame 35 of Mug; (b) corresponding depth map; and (c) synthesized view. The depth map is quantized to 256 gray levels where the depth is inversely related to the brightness. Note that depth has also been histogram equalized to show the contrast between the object and the surrounding background. Holes shown in red correspond to regions that become uncovered.*

To illustrate this synthesis procedure, consider frame 35 from the Mug sequence in Section 3 as shown in Figure 2 (a). Pairwise matching is performed between frame 35 and every one of its neighbors. The local depth maps are then combined to form Figure 2 (b). Figure 2 (c) is the result of warping every pixel according to its depth to synthesize a translated virtual camera. Notice how the motion parallax effect is preserved: Points closer to the camera appear to move more than those farther away. The red regions in the figure correspond to previously occluded points in the scene which become visible from the synthesized viewpoint. Note that there is insufficient information from a single reference intensity-depth pair to adequately fill in these regions. However, incorporating more information may reduce the size and number of these regions; this fact motivates the proposed representations in the following sections.

3 View Interpolation

It is clear from Section 2.3 that novel views of the scene may be synthesized quite accurately and easily from a single reference intensity-depth pair. Further improvements can be made by introducing a second or multiple reference pairs. Hence, our first proposed representation consists of employing multiple reference pairs. One may derive this representation using the techniques described in Section 2 in the following steps:

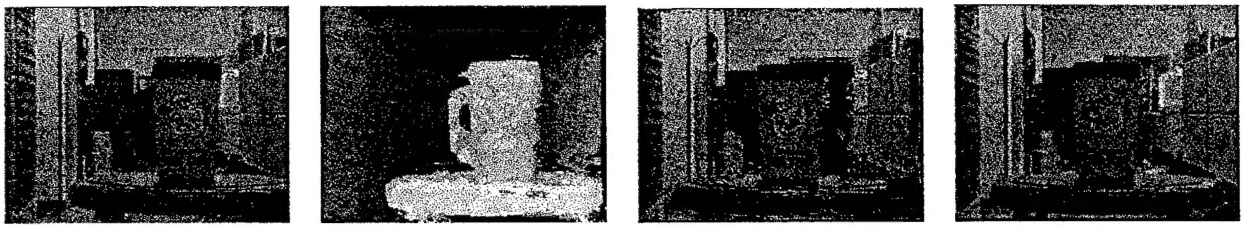
1. *Compute dense depth for every reference frame.*
   Assuming the reference frames have been chosen from the given image sequence(s), local

depth maps are obtained using the pairwise techniques of Section 2.1. The local depth maps are normalized and combined to form an accurate dense map [1].

2. *Estimate motion between reference frames.*
   For reference frames related by planar translation, it is sufficient to estimate the motion parameters (b_x, b_y) up to a scale factor using the least squares technique described in [5].

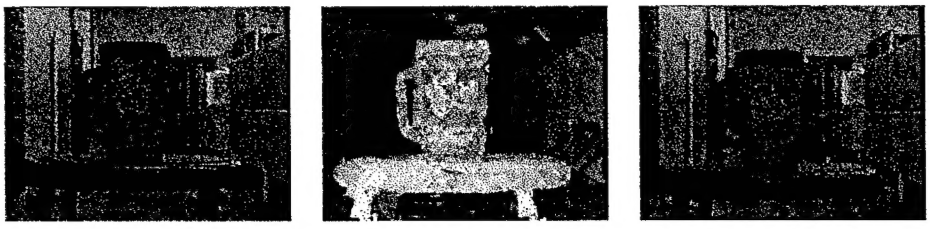
3. *Discard neighboring frames to form representation.*
   Notice that their use affects only the quality of the representation and not its compactness. The representation consists of the depth and intensity maps corresponding to the reference frames.

4. *Generate view estimates and combine to form desired view.*
   To synthesize a desired view, the transformation in Section 2.3 may be separately applied to each reference intensity-depth pair to create view estimates. Information from one view estimate will likely fill in the holes of another view estimate, and thus the overall number of holes may be reduced by combining view estimates together. Hence, multiple reference pairs help to overcome problems of occlusion during synthesis.



(a) (b) (c) (d)

Figure 3: *Reconstruction of horizontal view from reference frame 35 and 65 of Mug: (a) intensity image frame 65 of Mug; (b) corresponding depth map; (c) view estimate using only reference frame 65; and (d) reconstructed view combining view estimates.*



(a) (b) (c)

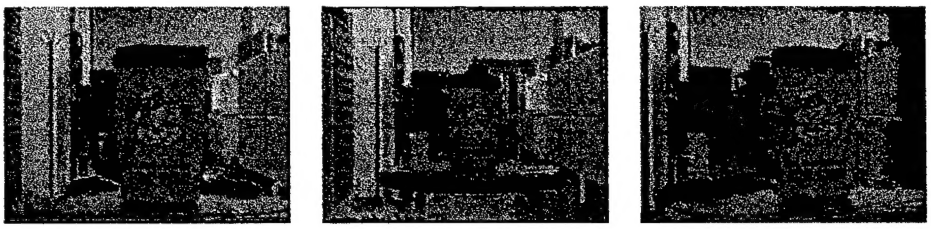
Figure 4: *Reconstruction of vertical view from reference frame 35 of Mug and frame 37 of Mug2: (a) intensity image frame 37 of Mug2; (b) corresponding depth map; and (c) reconstructed view.*

The above steps are applied to a real-world scene filmed by a camcorder undergoing unknown horizontal translation at two different elevations. The two sequences, known as Mug and Mug2, were digitized to 320×240 and subsampled temporally to obtain eighteen Mug frames and seven

Mug2 frames. Three frames, frames 35 and 65 from Mug and frame 37 from Mug2, were chosen to serve as reference frames; Figures 2, 3, and 4 show these reference pairs, respectively.

Using reference frames 35 and 65, the midpoint view along the same horizontal trajectory is chosen to be reconstructed. Using only reference frame 35 or 65 leads to the view estimates shown in Figure 2 (c) and Figure 3 (c), respectively. Since the holes in the view estimates do not overlap, one would expect improved results after combining the view estimates. As shown in Figure 3 (d), the combined result quality is good for the most part. The horizontal edges, *e.g.* top of the door, top of the mug, specularities in front of the stool, and the drawers, have been reconstructed quite well. The proposed approach takes care of problems in occluded regions; there are only a few errors to the right of the mug and near the mug handle. These artifacts arise because the depth edges were not localized perfectly.

To generate a view not originally scanned by the camcorder, reference frames 35 and 37 are used to synthesize the midpoint on the vertical trajectory relating the two views; the result is given in Figure 4 (c). The image is a reasonable estimate of the desired view. As before, the most troublesome region in the image lies inside the handle of the mug.



(a) (b) (c)

Figure 5: *Examples of synthesized views using multiple reference frames: (a) translation toward scene; (b) translation away from scene; and (c) arbitrary rotation and translation.*

More interesting views not necessarily confined to the x-y plane may be reconstructed with this representation. For instance, the viewpoint of a camera translated toward the scene can also be rendered quite easily; the resulting image is given in Figure 5 (a). Note that this view differs from a simple "zoom-in" since the latter requires only a larger focal length and it does not uncover occluded regions. The two regions above the stool are marked red because none of the reference frames has information about what lies behind the stool in the scene. Figure 5 (b) shows the view translated away from the scene with the uncovered regions marked accordingly. Finally, Figure 5 (c) shows an oblique view of the scene taken by rotating the camera 10° clockwise and translating along both the x and z axes. The quality of the synthesized image is quite good given the amount of uncovered regions.

4 Multivalued Representation

In representing a 3-D scene, it is common for the images to be very similar and to exhibit a lot of redundancy. This fact is especially true when the images come from arbitrary translational motion in the x-y plane since the depth of scene points remains fixed in all the images. One possible compact representation for this case would involve remapping all visible information with respect to one particular frame. We thus consider exploiting the redundancy to form a multivalued representation (MVR) of depth and intensity. The MVR separates information into levels of occlusion and can easily handle points occluded from reference viewpoint.

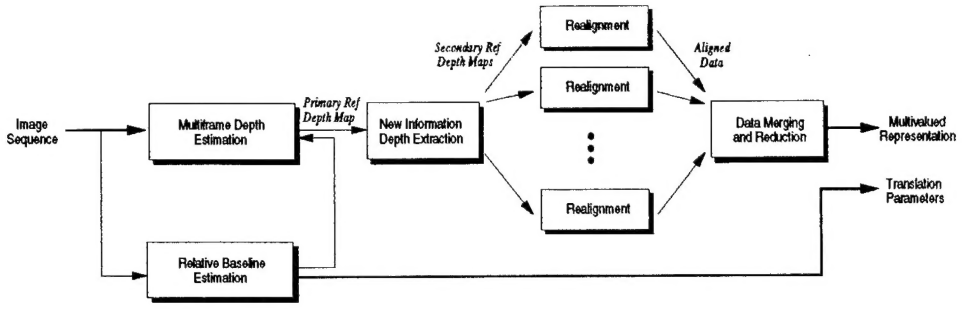


Figure 6: *Block diagram for the multivalued representation.*

To build a MVR from a set of images, one first selects a single frame, denoted as the *primary reference frame* or PRF, for which the representation is defined. As diagrammed in Figure 6, the following steps are then performed:

1. *Estimate motion parameters between PRF and each neighbor.*
   Point correspondences are first established by the approach of Zhang, Deriche, et al. [16]. Since only relative motion may be estimated, one horizontal parameter is fixed to 1. The remaining motion parameters are estimated using least squares [5].

2. *Calculate dense depth for PRF using multiframe algorithm.*
   The depth map corresponding to the PRF is computed using all neighboring frames simultaneously as described in Section 2.2.

3. *Compute depth for new information in other frames.*
   The PRF depth may then be warped to each of its neighbors' coordinate system. The points that are unmapped in each frame correspond precisely to the "new" or previously occluded information in that frame compared to the PRF. A multiframe algorithm is then executed to estimate depth for all the new information regions. Notice that this method is faster than calculating dense depth with respect to every frame.

4. *Fit piecewise 3-D surfaces through depth maps.*
   Clustering techniques are performed to segment regions of support in the depth domain [14]. Noisy depth estimates are then smoothed by fitting piecewise surfaces over the appropriate regions of support [11].

5. *Merge and reduce data to produce final MVR.*
   All of the new depth and intensity information are compensated back to the PRF by applying the inverse baselines. The depth map corresponding to the PRF is assigned to the first level of the MVR (level 0). Compensated points are added to the appropriate level by comparing existing points.

The final result consists of a multivalued array of intensities and depths corresponding to the primary reference frame. Notice that the information contained in the MVR consists of the union of intensity and depth that can be extracted from the original image data.

As before, we consider the Mug and Mug2 sequences, where only nine frames of Mug and four from Mug2 are used. Frame 50, shown in Figure 7 (a), is selected as the primary reference frame for the representation. Using the multiframe algorithm of Section 2.2 leads to the depth map found in Figure 7 (b). Notice the accuracy of the estimated depths especially the descending walls. The synthesis techniques of Section 2.3 may be applied to this depth map to obtain an estimate of, say, frame 21. If this view estimate is compared with the original image (see Figure 7 (c)), one

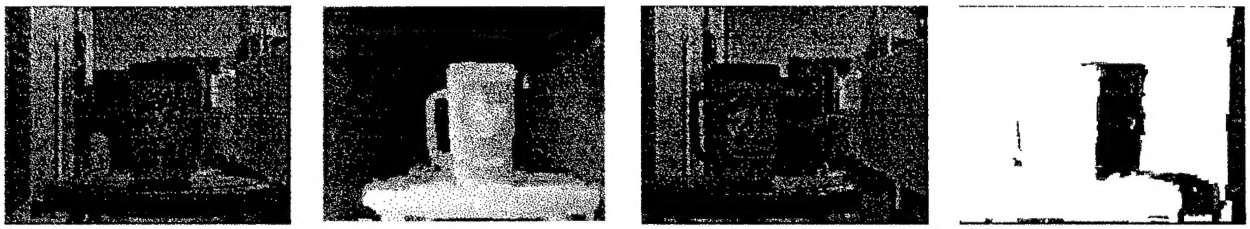
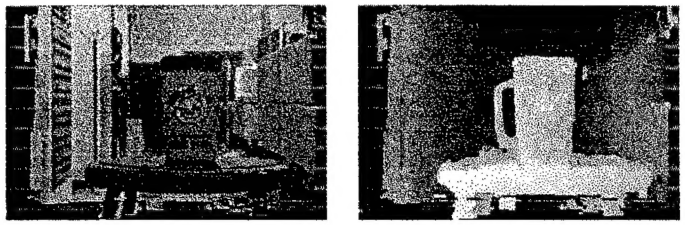


Figure 7: *Example of estimating new information: (a) intensity PRF 50 of Mug; (b) depth PRF 50; (c) intensity frame 21 of Mug; and (d) new information in frame 21 wrt frame 50. As expected, the algorithm identifies the cubicle located behind the mug as well as the right border of the image, both obscured from view in frame 50.*

can easily extract the new information contained in frame 21 with respect to the PRF as shown in Figure 7 (d).

Applying the above algorithm, dense depth corresponding to the points visible from the PRF as well as points occluded in this frame are recovered. The intensity and depth information in level 0 are shown in Figures 8 (a) and (b). Points shown in blue correspond to regions without intensity and depth. The shape of the mug and the stool have been recovered quite well. Notice that the left and right sides descend in depth as expected. Also, the dimensions of the original image have been expanded and the points seen along the borders have been recovered. Even the legs of the stool have been extended.



(a) (b)

Figure 8: *Recovered information for level 0 of the MVR: (a) intensity and (b) depth. The depth is quantized to 256 gray levels where the depth is inversely related to the brightness. Note that depth has also been histogram equalized to show the contrast between the object and the surrounding background.*

Figures 9 (a) and (b) show the recovered information in the second level of the MVR. Most of the information corresponds to points that are located behind the mug. The cubicle and the wall are both recovered from behind the mug since they were seen in some of the original images. Moreover, the ground obscured by the stool is revealed in this level. By filling in points from level 0 as in Figure 10, it appears that the mug and most of the stool have been removed. Notice that the bottom portion of the legs and part of the stool remain since the regions behind them were occluded in the original images.

The reconstruction techniques of Section 2.3 are applied to generate the original images. As an example, Frame 21 has been reconstructed in Figure 11 (a). Notice that the reconstructed quality is quite good. Similar quality is obtained in the other reconstructed images as seen in Figures 11 (b)–(d). The average PSNR for reconstructed images is 30.707 dB.

(a) (b)

Figure 9: *Recovered information for level 1 of the MVR: (a) intensity and (b) depth. The cubicle located behind the mug was recovered in both intensity and depth domains. Also the wall behind the mug handle and the floor behind the stool are revealed.*

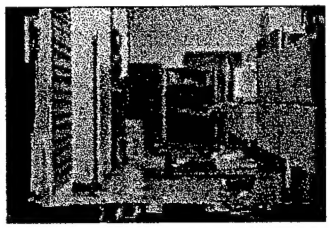


Figure 10: *Points from level 1 are combined with points from level 0 to put the representation in context.*

Synthesized views of the scene may be generated in a similar manner. Translations toward and away from the scene are given in Figures 12 (a) and (b), respectively. Figures 12 (c) and (d) show the virtual camera undergoing arbitrary motion. Despite an increase in the number of artifacts for these views, the resulting images are reasonable and provide a convincing sense of depth.

5 Discussion and Future Work

We have proposed two depth-based representations to address the problem of compact representation for image reconstruction and new view synthesis. The results from previous sections verify the effectiveness of both approaches. In the first case, multiple reference intensity-depth pairs serve as the representation. They provides an intuitive method for compressing and representing the information in a given image set while allowing for the generation of novel viewpoints of the scene.

(a) (b) (c) (d)

Figure 11: *Examples of reconstructed views using MVR: (a) frame 21; (a) frame 37; (a) frame 40; and (d) frame 80.*

(a) (b) (c) (d)

Figure 12: *Examples of synthesized views using MVR: (a) translation toward scene; (b) translation away from scene; (c) and (d) arbitrary rotation and translation.*

The use of multiple reference frames leads to improved results especially in uncovered regions.

The second representation redefines the available information with respect to a single multivalued array of depth and intensity. Because of its similarity to layered representations, it accumulates information seen in the union of frames as well as minimizes the redundancy of the overall representation. However, it can also lead to the generation of new views and handle more complex shapes than modeled by 2-D affine motion. Since it is primarily a depth-based representation, it is also capable of overcoming problems of occlusion during synthesis. A multiframe depth estimation algorithm has been shown to be effective in incorporating all frames simultaneously to estimate depth.

We are currently exploring several extensions to these depth-based representations. One obvious extension is to remove the planar translation restriction and allow arbitrary camera motion. We also plan to improve techniques for clustering and surface modeling, especially in the hope of further compacting the representation. Finally, we will examine issues involved in virtual environment applications (*e.g.* flyarounds and flythroughs).

References

[1] N. L. Chang, "View reconstruction from uncalibrated cameras for three-dimensional scenes," Master's thesis, University of California at Berkeley, 1994.

[2] N. L. Chang and A. Zakhor, "Arbitrary view generation for three-dimensional scenes from uncalibrated video cameras," in *Proceedings of ICASSP*, pp. 2455–2458, Detroit, MI, 8–12 May 1995.

[3] N. L. Chang and A. Zakhor, "Intermediate view reconstruction for three-dimensional scenes," in *Proceedings of ICDSP*, vol. 2, pp. 636–641, Nicosia, Cyprus, 14–16 July 1993.

[4] N. L. Chang and A. Zakhor, "Multivalued representations for image reconstruction and new view synthesis." In preparation.

[5] N. L. Chang and A. Zakhor, "View generation for three-dimensional scenes from video sequences," *IEEE Trans. on Image Proc.*, vol. 6, no. 4, Apr. 1997. To appear.

[6] S. E. Chen and L. Williams, "View interpolation for image synthesis," in *Proceedings of SIGGRAPH*, pp. 279–288, New York, NY, 1–6 Aug. 1993.

[7] T. Darrell and A. Pentland, "Robust estimation of a multi-layered motion representation," in *IEEE Workshop on Visual Motion*, pp. 173–178, Princeton, NJ, 7–9 Oct. 1991.

[8] J. D. Foley, A. van Dam, et al., *Introduction to Computer Graphics.* Addison-Wesley, 1994.

[9] T. Kanade, P. J. Narayanan, and P. W. Rander, "Virtualized reality: Concepts and early results," in *IEEE Workshop on Representation of Visual Scenes*, pp. 69–76, Cambridge, MA, June 24 1995.

[10] S. Laveau and O. Faugeras, "3-D scene representation as a collection of images and fundamental matrices," Tech. Rep. 2205, INRIA, Feb. 1994.

[11] H. Maître and W. Luo, "Using models to improve stereo reconstruction," *IEEE Trans. on Patt. Anal. Mach. Intell.*, vol. 14, no. 2, pp. 269–277, Feb. 1992.

[12] M. Okutomi and T. Kanade, "A multiple-baseline stereo," *IEEE Trans. on Patt. Anal. Mach. Intell.*, vol. 15, no. 4, pp. 353–363, Apr. 1993.

[13] H. S. Sawhney and S. Ayer, "Compact representations of videos through dominant and multiple motion estimation," *IEEE Trans. on Patt. Anal. Mach. Intell.*, vol. 18, no. 8, pp. 814–830, Aug. 1996.

[14] J. Y. A. Wang and E. H. Adelson, "Representing moving images with layers," *IEEE Trans. on Image Proc.*, vol. 3, no. 5, pp. 625–638, Sept. 1994.

[15] Y. Weiss and E. H. Adelson, "A unified mixture framework for motion segmentation: Incorporating spatial coherence and estimating the number of models," in *Proceedings of CVPR*, pp. 321–326, San Francisco, CA, 18–20 June 1996.

[16] Z. Zhang, R. Deriche, et al., "A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry," *Artificial Intelligence*, vol. 78, no. 1–2, pp. 87–119, Oct. 1995.